用于处理不努力作答的标准化残差系列方法和混合多层模型法的比较*

刘　玥[1]　刘红云[2,3]

([1]四川师范大学脑与心理科学研究院，成都　610066)

([2]应用实验心理北京市重点实验室) ([3]北京师范大学心理学部，北京　100875)

摘要

文章采用模拟研究，分别在混合多层模型假设满足和违背的情境下，比较了混合多层模型方法与标准化残差系列方法在识别不努力作答和参数估计方面的表现。结果显示：（1）不存在不努力作答或其严重性低时，各方法表现接近；（2）不努力作答严重性高时，固定参数迭代标准化残差法普遍更优，混合多层模型法仅在假设满足且两种作答反应时差异大的条件下表现较好。建议实际应用中优先选择固定参数迭代标准化残差法。

关键词　不努力作答，标准化反应时残差，迭代净化，混合多层模型，贝叶斯估计

1 引言

在对学生的人格、技能和能力等潜在特质进行测量时，最主要的目的是基于测验信息得到学生潜在特质的有效估计值。然而，在实际中，难免有学生在测验时作出不努力作答（non-effortful response），为测验带来与结构无关的污染。总的来说，不努力作答具有反应时短、正确率低、提供的心理测量学信息少三个特征（Wise, 2015; 2017）。测验中出现的不努力作答会对测验信效度造成各种不利影响。首先，很多情况下被试的能力值会被低估（Rios et al., 2017; Wise, 2015; Wise & DeMars, 2006; Wise & Kingsbury, 2016），进而造成群组分数的差异（Borghans & Schils, 2012）。其次，题目参数估计值的偏差会增大（Wise & DeMars, 2006）。第三，如果不同子群体中不努力作答的比例不同，这种差异还可能导致项目功能差异（Setzer et al., 2013）。第四，测验的信息量、信度会出现偏差（Wise & DeMars, 2006）。第五，测验

* 收稿日期：2021-04-08
国家自然科学基金项目（32071091）
通讯作者：刘红云，E-mail：hyliu@bnu.edu.cn

所测量的结构也可能变化，会聚效度出现偏差（Wise & DeMars, 2006）。最后，与测验有关的预测变量和结果变量之间的关系，假设检验得到的结论等，都可能出现偏差（Clark et al., 2003）。因此，在测验（特别是低利害情境下测验）的数据分析中，有必要通过科学的方法，处理不努力作答，减小其不利影响，得到更准确的参数估计结果。

不努力作答的处理主要包括识别并降低权重和在模型中处理两种思路。识别并降低权重是指在数据清理时首先识别不努力作答，再在数据分析时降低其权重（Ranger et al., 2019; Rios et al., 2017）。降低权重部分最极端和常用的方式是替换为缺失（e.g., Köhler et al., 2017; Rose, 2013）。识别部分较经典的方法是标准化残差法。该方法将观测反应时与其理论分布比较，以识别反应时异常短的不努力作答（Qian et al., 2016）。标准化残差法的优势在于背后有特定的理论模型（分布），不需要通过观察设定阈值，也不存在无法找到阈值的特例，可以自动化大批量应用。此外，van der Linden 和 Guo（2008）曾提出贝叶斯残差法，将反应时观测值与基于作答反应和反应时计算的后验预测密度比较，以识别不努力作答。该方法与标准化残差法都面临着参数污染严重时表现差的问题（Wang, Xu, Shang, & Kuncel, 2018）。最近，针对这一缺陷，Liu 和 Liu（2021）采用筛选努力作答群体估计题目参数，固定题目参数并迭代净化的策略改进标准化残差法的表现，提出了固定参数迭代标准化残差法，并得到了较好的效果。

在模型中处理主要指使用混合模型，区分努力作答和不努力作答的数据，并分别采用不同的模型拟合（Molenaar et al., 2018; Wang & Xu, 2015; Wang, Xu, & Shang., 2018; Wise & DeMars, 2006）。与识别并降低权重的两阶段方法相比，混合模型能够一次性解决不努力作答识别及参数估计的问题。并且，贝叶斯估计的马尔科夫链蒙特卡洛（Markov Chain Monte Carlo, MCMC）算法的发展，较好地解决了这类模型参数估计的问题。混合模型中最有代表性的方法是由 Wang 和 Xu（2015）基于 van der Linden（2007）的多层模型提出的混合多层模型（mixture hierarchical model，MHM）。它的主要思想是根据两种作答行为的特点，对总体的作答反应模型和反应时模型进行分解。模拟研究证明，当数据中同时含有努力作答与不努力作答时，MHM 相比于传统多层模型能够得到更准确的参数估计结果（Wang & Xu, 2015）。Wang，Xu，Shang 和 Kuncel（2018）还采用模拟研究，对贝叶斯残差法和 MHM 进行了比较。结果表明 MHM 在正确识别率和错误拒绝率上表现都较好，特别是当异常作答的比例较高时，该模型的优势更加明显。后来的研究者在 MHM 基础上又进行了一系列拓展研究（Lu et al., 2020; Ulitzsch et al., 2020; Wang, Xu, & Shang., 2018）。总的来说，MHM 最大的优势在于能够同时完成异常反应的识别和模型参数的估计。但是，该方法主要有三个局限性：一是

包含关于不努力作答的正确率和反应时分布的强假设，如果不满足，可能无法得到准确的识别结果；二是不努力作答比例较低时容易出现问题，例如，当不努力作答的比例较小或者样本量较小时，有时很难得到收敛的结果（Ranger et al., 2019）；三是计算复杂耗时长。总的来说，关于混合模型的研究基本上都以 Wang 和 Xu（2015）的混合多层模型为基础展开，因此，本研究也关注该模型和标准化残差系列方法的比较。

尽管标准化残差法和混合多层模型法作为两种处理思路的代表，具有不同的优缺点和适用条件，但是目前对这两类方法进行系统比较的模拟和应用研究仍较少，且选用的残差法没能反映该方法最新的研究进展（Liu & Liu, 2021）。虽然 Wang，Xu，Shang 和 Kuncel（2018）的研究对贝叶斯残差法和混合多层模型法进行了比较。但是，该研究也存在一些局限性。首先，研究中设置的基于残差模型产生数据的情境，仅违背了混合多层模型中关于反应时模型的假设，而不努力作答的答对概率仍符合其假设，因此并不能算作反应时和作答反应均违背其假设的情况。其次，贝叶斯残差法本身计算较复杂且在实际中很少应用，此外，该方法与标准化残差法同样面临在数据污染严重情况下表现差的问题。新的固定参数迭代标准化残差法（Liu & Liu, 2021）相对于贝叶斯残差法计算和原理都更为简单，其能否弥补传统方法的缺陷，得到与 MHM 相近甚至更好的结果？由于固定参数迭代标准化残差法相比 MHM 前提假设较少，其是否具有更好的稳健性，也是方法的理论和实践研究关注的焦点。目前，尚未有研究系统比较标准化残差系列方法和混合多层模型法。因此，两类方法在不同条件下的表现和效率，是本研究关注的主要问题。

本文首先回顾了 3 种标准化残差法和 Wang 和 Xu（2015）的混合多层模型法，然后分别构造了产生数据完全符合、反应时和作答反应均不符合混合多层模型假设的两种情境。采用模拟研究的方法，在不同条件下对两类方法识别和参数估计结果的准确性进行比较，以期能够对各方法的优缺点和适用范围有更深入的认识，为实际应用者提供方法选择的建议。另外，研究还将两类方法应用于一项测验的实际数据，对模拟研究的结果进行了进一步印证与补充。

2 标准化残差系列方法

原始标准化残差法（original standard residual method，OSR）首先基于原始数据，应用 van der Linden （2007）的多层模型估计参数。该模型包括两个水平，第一水平是测量模型，包括作答反应部分的 IRT（Item response theory）模型：

$$P(Y_{ij}=1|\theta_i)=\frac{exp\,(a_j(\theta_i-b_j))}{1+exp\,(a_j(\theta_i-b_j))}, \tag{1}$$

和反应时部分的标准对数正态分布模型：

$$ln(t_{ij})|\tau_i \sim N(\beta_j-\tau_i,\alpha_j^{-2}), \tag{2}$$

其中，$P(Y_{ij}=1|\theta_i)$表示被试 i（i = 1,…, I）在题目 j（j = 1,…, J）上正确作答的概率，t_{ij}表示被试 i 在题目 j 上的反应时，a_j和b_j分别是题目 j 的区分度参数和难度参数，β_j表示题目 j 的时间密度参数，α_j表示题目 j 的时间区分度参数，$N(.)$表示正态分布，θ_i和τ_i是被试 i 的能力参数和速度参数。

在第二水平（个体水平），假设被试参数(θ_i,τ_i)服从二元正态分布，能力和速度参数的均值为(μ_θ,μ_τ)，能力参数的方差为σ_θ^2，速度参数的方差为σ_τ^2，能力和速度参数的协方差为$\sigma_{\theta\tau}$。

然后，计算被试 i 在题目 j 上的标准化反应时残差

$$\hat{e}_{ij}=\hat{\alpha}_j\left(ln(t_{ij})-(\hat{\beta}_j-\hat{\tau}_i)\right). \tag{3}$$

根据标准化反应时残差服从标准正态分布$e_{ij}\sim N(0,1)$进行判断，基于显著性水平为 0.05 的标准正态分布左侧检验，如果$\hat{e}_{ij}$ <–1.645，则认为被试 i 在题目 j 上是不努力作答（Qian et al., 2016）。

当数据污染严重时，为改进 OSR 对题目参数估计不准确的问题，固定参数标准化残差法（conditional estimate standard residual, CSR）建议首先通过混合模型，筛选努力作答群体，并基于该群体获得较准确的题目参数估计结果。然后将题目参数固定，对被试参数进行条件估计。最后，基于这些参数估计结果，应用 OSR 识别不努力作答（Liu et al., 2020）。

固定参数迭代标准化残差法（conditional estimate with fixed item parameters standard residual method using iterative purifying procedure，CSRI）在 CSR 的基础上不断应用迭代净化过程，提高被试参数估计准确性，以适用于数据污染严重的情况（Liu & Liu, 2021）。

在使用 OSR，CSR 和 CSRI 之后，需将识别出的不努力作答记为缺失，基于 van der Linden （2007）的多层模型重新估计所有参数值。

3 混合多层模型法

混合多层模型（MHM）根据努力作答和不努力作答的特点，对总体的作答反应模型和反应时模型作分解（Wang & Xu, 2015）。

在作答反应模型部分，假设被试 i 在题目 j 上答对的概率为

$$P(Y_{ij}=1|\Delta_{\mathrm{ij}})=(1-\Delta_{\mathrm{ij}})P(Y_{ij}=1|\Delta_{\mathrm{ij}}=0)+\Delta_{\mathrm{ij}}P(Y_{ij}=1|\Delta_{\mathrm{ij}}=1), \tag{4}$$

其中，Δ_{ij}是表示作答行为分类的潜变量，$\Delta_{ij}=1$表示被试 i 回答题目 j 是不努力作答，$\Delta_{ij}=0$表示是努力作答。如果$\Delta_{ij}=0$，可使用两参数 logistic（2PL）模型预测努力作答的答对概率（见公式（1））。如果被试 i 回答题目 j 是不努力作答（$\Delta_{ij}=1$），则答对概率是g_j。即

$$P(Y_{ij}=1|\Delta_{ij}=1)=g_j. \tag{5}$$

在反应时模型部分，假设对于被试 i 和题目 j，观察到的反应时T_{ij}^{obs}可以表示为

$$T_{ij}^{obs}=(1-\Delta_{\mathrm{ij}})T_{ij}+\Delta_{\mathrm{ij}}C_{ij}, \tag{6}$$

其中，T_{ij}表示被试 i 努力作答题目 j 所需要的时间，C_{ij}表示被试 i 不努力作答题目 j 所需要的时间。努力作答的反应时服从对数正态分布（见公式（2））。假定不努力作答的反应时也服从对数正态分布

$$\ln(C_{ij})\sim\mathrm{N}(\mu_c,\sigma_c^2), \tag{7}$$

其中，μ_c表示不努力作答反应时对数正态分布的均值，σ_c^2表示分布的方差。

在实际中，不努力作答部分模型所包含的强假设可能会遭到违背。具体表现在，第一，该模型假设异常作答的正确率为g_j，即所有被试在同一道题上不努力作答的答对概率是相同的。但是 Feinberg 和 Jurich（2018）发现，不同能力水平被试在相同题目上不努力作答的正确率不同。第二，该模型假设不努力作答行为的反应时服从均值和标准差恒定的对数正态分布。然而实际中不努力作答的反应时可能和被试因素（例如，学业能力、作答速度等），或者题目因素（例如，题目位置，题型等）相关（e.g., Molenaar et al., 2018）。

4 研究一：标准化残差系列方法与混合多层模型法比较的模拟研究

4.1 研究方法

4.1.1 研究设计

模拟研究共含两种情境。情境 1，数据符合混合多层模型假设；情境 2，不努力作答的反应时和作答反应均不符合混合多层模型假设。每种情境都采用混合实验设计，组内变量为 OSR，CSR，CSRI 和 MHM。

对于情境 1，组间变量有三个：（1）不努力作答规模（π，含有不努力作答的被试所占比例）：0%，20%，40%；（2）不努力作答严重性（π_i^{non}，含有不努力作答被试的不努力作答题目比例）：低（$\pi_i^{non}\sim U(0,0.25)$），高（$\pi_i^{non}\sim U(0.5,0.75)$）；（3）两种作答反应时差异

（d_{RT}，不努力作答与努力作答的反应时差异）：小，大。不努力作答规模$\pi = 0\%$表示所有被试在所有题目上均努力作答，设置该水平是为了考察在没有不努力作答的条件下，各方法可能存在的超识别问题。根据π和π_i^{non}的组合，生成数据中不努力作答的比例覆盖了 0%，2.5%，5%，12.5% 和 25%几种情况。组间变量共形成 2×2×2+1 = 9 种实验条件。

对于情境 2，由于不努力作答的反应时基于残差模型生成，无法从整体上控制两种作答反应时均值的差异，因此不考虑d_{RT}。另外，$\pi = 0\%$的数据产生方式与情境 1 完全相同。因此，情境 2 中考虑的组间变量包括：（1）π：20%，40%； π_i^{non}：低，高。组间变量共形成 2×2 = 4 种实验条件。

参照前人研究，模拟研究的样本容量固定为 2000，题目数固定为 30（Wang & Xu, 2015; Wang, Xu, & Shang., 2018; Wang, Xu, Shang, & Kuncel, 2018）。

4.1.2 数据生成

题目参数产生值的分布为$a_j \sim U(1,2.5)$， $b_j \sim N(0,1)$，$\alpha_j \sim U(1.5,2.5)$，$\beta_j \sim U(-0.2,0.2)$。这些分布的选择保证了产生的作答反应和反应时与真实数据类似（van der Linden, 2007; Wang & Xu, 2015; Wang, Xu, Shang, & Kuncel, 2018）。被试参数（θ_i, τ_i）产生于二元正态分布，两个参数的均值都是 0，方差分别为 1 和 0.25，协方差为 0.25。采用这种方式，能够保证θ_i和τ_i的相关固定为中等水平，即高能力被试倾向于作答速度较快（Wang & Xu, 2015; Wang, Xu, & Shang., 2018; Wang, Xu, Shang, & Kuncel, 2018）。下面分不同情境介绍数据生成的具体方式。

（1）情境 1

首先，利用题目参数和被试参数的真值，基于 van der Linden（2007）的多层模型模拟生成努力作答的作答反应和反应时。然后生成不努力作答数据，包含以下步骤：（a）基于π选出相应数量的被试。因为速度较慢的被试倾向于猜测作答（不努力作答），因此，从真实速度 τ_i 最低 33%的被试中随机选择 60%的被试，中间 34%的被试中随机选择 30%的被试，最高 33%的被试中随机选择 10%的被试，作为含有不努力作答的被试群体（Wang, Xu, & Shang., 2018）；（b）由于不努力作答可能随机发生在任何题目上（Pastor et al., 2019），根据π_i^{non}，对于π中的被试随机选择相应数量的不努力作答（Wang, Xu, & Shang., 2018）；（c）对所有不努力作答，参考 Wang 和 Xu（2015），将答对概率（g_j）均设定为 0.25，模拟产生作答反应；按照取自然对数后的反应时服从正态分布$\mathrm{N}(\mu_c, \sigma_c^2)$ 模拟产生反应时（Liu & Liu,

2021），对两种作答反应时差异小和大两种情况，不努力作答反应时取对数后的分布分别服从 $N(-1,0.5^2)$和 $N(-2,0.5^2)$。最后，使用不努力作答的作答反应和反应时替换原有数据中相应位置的数据。

（2）情境 2

情境 2 和情境 1 的区别在于生成不努力作答数据的方式。对于作答反应，基于 Feinberg 和 Jurich（2018）的发现，不同能力水平被试快速猜测的正确率不同，能力高的被试正确率高于能力低的被试。因此，按能力值将被试分为 3 组，分别为能力值小于-0.44，能力值介于-0.44 到 0.44 之间，能力值大于 0.44（每组被试约占 1/3）,每组对应不努力作答的答对概率分别为 0, 0.25 和 0.5。因此，情境 2 不符合混合多层模型关于不同被试不努力做答答题概率相同的假设。产生不努力作答反应时的步骤为（Wang, Xu, Shang, & Kuncel, 2018)：（a）基于反应时服从对数正态分布的假设，利用时间密度参数、时间区分度参数的真值，对于速度为 0 的被试，计算每道题目反应时取自然对数后最低 5%的临界值（e.g., 对于题目 j 为$P^{0.05}_{\ln(t_j)}$)；（b）对于题目 j，在$U\left(\exp(-5),\exp\left(P^{0.05}_{\ln(t_j)}\right)\right)$的区间内随机取一个值作为不努力作答的反应时。此时不努力作答的反应时符合残差模型，可以被看作整个反应时分布中的异常值。但是，与情境 1 不同，它们的分布不满足对数正态分布，因此不符合 MHM 的假设。

采用蒙特卡洛模拟研究的方法，使用 R 软件（R Development Core Team, 2009）产生两种情境不同条件下的作答反应和反应时数据，每种条件下数据重复模拟 30 次（e.g., Lu et al., 2020; Wang, Xu, Shang, & Kuncel, 2018）。

4.1.3 参数估计

参考前人研究（Lu et al., 2020; Wang & Xu, 2015; Wang, Xu, & Shang, 2018; Wang, Xu, Shang, & Kuncel, 2018），研究应用贝叶斯框架下基于 Gibbs 抽样的 MCMC 算法估计参数后验分布，进而计算后验均值得到参数的点估计值。这一过程利用 JAGS4.3.0 软件自编语句实现（Plummer, 2003）。

对于 MHM，先验分布的设置参考了前人研究（Wang, Xu, & Shang, 2018; Wang, Xu, Shang, & Kuncel, 2018）。努力作答部分题目参数的先验分布为：$a_j \sim lognormal(0,1), b_j \sim N(0,1), \alpha_j^2 \sim lognormal(0,1), \beta_j \sim N(0,1)$；不努力作答部分题目参数的先验分布为$g_j \sim beta(2,10), \mu_c \sim N(-3,0.1), \sigma_c^2 \sim \mathrm{Inv}-\gamma(10,0.1)$；被试参数采用与产生值相同的分布。对于 OSR，CSR 和 CSRI，应用 van der Linden（2007）的多层模型时先验分布的设

置与 MHM 中努力作答部分模型参数的先验分布一致。迭代的初始值在每个参数先验分布中随机抽取样本得到。经过前期试验得到正式研究的 MCMC 迭代设置参数。MCMC 链条数量固定为 2，每条链的迭代次数为 10000，前面 5000 次作为 burn-in，thinning rate 固定为 5。由于 MHM 较为复杂，试验发现在原有设置基础上即使增加迭代次数，收敛情况也不会有明显改变，且每次估计时间已长达 9 小时，因此出于估计效率的角度，迭代设置参数仍保持原有设置。采用 PSRF<1.1 作为判断每条链收敛的标准（Gelman & Rubin, 1992; Matzke et al., 2017）。

4.1.4 评价标准

研究从三个方面对两种不同类型的方法进行比较。

（1）收敛情况

根据 PSRF 指标，统计了各方法下各参数估计的收敛比例。

（2）识别准确性

评价识别准确性的指标分为基于不努力作答的正确识别率（true positive rate, TPR）和错误识别率（false discovery error, FDR）。TPR 是指正确识别的不努力作答占真正不努力作答的比例，FDR 是指错误识别的不努力作答（即真正的努力作答）占所有识别出的不努力作答的比例。由于研究目的是识别不努力作答，因此，TPR 越高，说明识别出的不努力作答越全，越有利于得到准确的参数估计结果。基于这一目的，在评价识别准确性时，以 TPR 为主要依据。另外，当模拟数据中不存在不努力作答时（$\pi = 0\%$），无法计算 TPR，而 FDR 始终为 1，在这种情况下计算误检率（false positive rate, FPR），即错误识别出的不努力作答占所有努力作答的比例，类似于第 I 类错误概率。最后，计算了各方法在各条件下识别出的不努力作答占所有作答的比例（proportion, Pr）。

（3）参数估计结果准确性

研究使用偏差（bias）和误差均方根（RMSE）评价参数估计的返真性，计算公式如下

$$bias = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{H}\sum_{h=1}^{H}\left(o_h - \hat{o}_h^{(l)}\right), \quad (8)$$

$$RMSE = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\frac{1}{H}\sum_{h=1}^{H}\left(o_h - \hat{o}_h^{(l)}\right)^2}, \quad (9)$$

其中，o_h表示参数真值，$\hat{o}_h^{(l)}$表示第 l 次重复中参数的估计值（对于题目参数 h=j，对于被试参数 h=i），H 表示题目数量（H=J）或者被试数量（H=I），L=30 表示每种条件下的重复次

数。

4.2 模拟研究结果

4.2.1 参数收敛结果

OSR，CSR 和 CSRI 在所有重复中所有参数全部收敛。MHM 存在一定程度的不收敛问题。各条件下 MHM 不收敛的比例如表 1 所示。

表 1 各条件下 MHM 不收敛百分比（%）

	π	π_i^{non}	d_{RT}	作答分类参数（Δ_{ij}）	题目参数	被试参数	合计
	0%			0.05	0.00	0.00	0.05
情境 1	20%	低	小	15.83	0.00	0.00	14.80
			大	11.70	0.00	0.00	10.94
		高	小	11.10	0.00	0.00	10.38
			大	12.11	0.00	0.01	11.33
	40%	低	小	12.88	0.00	0.00	12.04
			大	12.73	0.00	0.00	11.91
		高	小	9.30	0.00	0.00	8.70
			大	13.15	0.00	0.00	12.30
情境 2	20%	低		16.75	0.00	0.00	15.67
		高		15.53	0.00	0.00	14.52
	40%	低		7.08	0.00	0.00	6.62
		高		11.93	0.00	0.00	11.15

注：π表示不努力作答规模，π_i^{non}表示不努力作答严重性，d_{RT}表示两种作答反应时差异，后同。合计是指不收敛参数占所有估计参数的百分比。

从表中可以看出，只有作答分类参数存在不收敛问题。其中，在全部努力作答的条件下，不收敛比例最低，为 0.05%，在π为 20%，π_i^{non}低的条件下，不收敛比例最高，为 14.80%（情境 1，d_{RT}小）和 15.67%（情境 2）。整体来看，情境 2 中不收敛百分比要大于情境 1 的情况。

以下的识别准确性和参数估计结果准确性评价指标仅针对所有收敛参数计算。

4.2.2 识别准确性结果

表 2 呈现了不同条件下各方法的识别准确性结果的均值。从表中可以看出，当不含有不努力作答时，MHM 估计得到的作答分类参数均为 1 个类别，而标准化残差系列方法的 FDR 约为 5%。从 TPR 来看，几乎所有条件下 CSRI 的 TPR 都大于 MHM。在大部分条件下，MHM 的 TPR 均最低。π_i^{non}越高，d_{RT}越大，CSRI 相对其他残差法的优势越大，MHM 的表现也越来越好。总的来看，情境 2 中 MHM 在 TPR 上的均值小于情境 1，标准化残差系列方

法则相对稳定。各条件下 MHM 的 FDR 均最小，CSRI 的 FDR 均最大。MHM 所识别出的不努力作答的比例大多小于 CSRI，这也反映在该方法呈现出较低的 TPR 和 FDR。

表　2 各条件下各方法识别准确性指标结果

情境	π	π_i^{non}	d_{RT}	指标	OSR	CSR	CSRI	MHM
情境 1	0%			FPR	0.05	0.05	0.06	0.00
	20%	低 (0.025)	小	TPR	0.59	0.59	**0.69**	0.39
				FDR	0.69	0.69	0.71	0.20
				Pr	0.05	0.05	0.06	0.01
			大	TPR	0.91	0.91	**0.97**	0.87
				FDR	0.47	0.49	0.53	0.09
				Pr	0.04	0.04	0.05	0.02
		高 (0.125)	小	TPR	0.19	0.25	**0.50**	0.03
				FDR	0.48	0.54	0.43	0.08
				Pr	0.04	0.07	0.11	0.00
			大	TPR	0.31	0.50	**0.93**	0.82
				FDR	0.16	0.36	0.28	0.07
				Pr	0.05	0.10	0.16	0.11
	40%	低 (0.050)	小	TPR	0.55	0.55	**0.65**	0.51
				FDR	0.46	0.45	0.47	0.20
				Pr	0.05	0.05	0.06	0.03
			大	TPR	0.87	0.87	**0.94**	0.91
				FDR	0.17	0.16	0.18	0.09
				Pr	0.05	0.05	0.06	0.05
		高 (0.250)	小	TPR	0.13	0.24	**0.49**	0.16
				FDR	0.23	0.31	0.23	0.10
				Pr	0.04	0.09	0.16	0.05
			大	TPR	0.17	0.49	0.93	**0.94**
				FDR	0.03	0.17	0.14	0.07
				Pr	0.04	0.15	0.27	0.25
情境 2	20%	低 (0.025)		TPR	0.77	0.78	**0.90**	0.64
				FDR	0.52	0.53	0.55	0.10
				Pr	0.04	0.04	0.05	0.02
		高 (0.125)		TPR	0.27	0.34	**0.72**	0.18
				FDR	0.17	0.35	0.24	0.01
				Pr	0.04	0.07	0.12	0.02
	40%	低 (0.050)		TPR	0.70	0.69	**0.82**	0.73
				FDR	0.22	0.21	0.22	0.11
				Pr	0.04	0.04	0.05	0.04
		高 (0.250)		TPR	0.20	0.29	**0.56**	0.13
				FDR	0.02	0.10	0.06	0.00
				Pr	0.05	0.08	0.15	0.03

注：TPR 表示正确识别率，FDR 表示错误识别率，FPR 表示误检率，Pr 表示识别出的不努力作答占所有作答的比例。π_i^{non}一列中括号内数字表示真实不努力作答的百分比。加粗的结果表示每种条件下 TPR 最高的结果。

4.2.3 参数估计结果

表 3、表 4 和表 5 分别展示了情境 1、情境 2 中各条件下各方法得到的参数估计准确性结果。从表中看出，当不含有不努力作答时，MHM 得到的各参数估计值 RMSE 普遍较小，标准化残差系列方法除了高估时间区分度参数外，其他参数估计值 RMSE 也较小。在π_i^{non}低的条件下，各方法得到的参数估计结果准确性差异不大，但是，除了π=40%，d_{RT}大的条件下两类方法得到的时间区分度参数 RMSE 较为接近，其余条件下 MHM 得到的时间区分度参数 RMSE 均明显小于标准化残差系列方法。总的来说，方法之间的差异主要体现在π_i^{non}高的条件下，当d_{RT}小时，CSRI 得到的参数估计结果准确性具有一定优势，当d_{RT}大时，CSRI 和 MHM 得到的参数估计结果准确性都具有更加明显的优势。

表 3 情境 1 中不含不努力作答条件下各方法参数估计准确性

评价标准	方法	OSR	CSR	CSRI	MHM
bias	a	–0.01	–0.01	–0.01	0.01
	b	0.00	0.00	0.00	0.00
	α	–0.21	–0.22	–0.26	0.00
	β	–0.07	–0.07	–0.08	0.02
	θ	0.00	–0.01	–0.01	0.01
	τ	–0.01	–0.01	–0.01	0.02
RMSE	a	0.11	0.11	0.11	0.10
	b	0.05	0.05	0.05	0.05
	α	0.22	0.22	0.27	0.03
	β	0.07	0.07	0.08	0.02
	θ	0.29	0.29	0.29	0.28
	τ	0.10	0.10	0.11	0.09

注：bias 表示偏差，RMSE 表示误差均方根，后同。

表 4 情境 1 中含有不努力作答条件下各方法参数估计准确性

π	评价标准	π_i^{non}	低								高							
		d_{RT}	小				大				小				大			
		方法	OSR	CSR	CSRI	MHM	OSR	CSR	CSRI	MHM	OSR	CSR	CSRI	MHM	OSR	CSR	CSRI	MHM
20%	bias	a	0.05	0.05	0.04	–0.03	0.01	0.01	0.00	–0.02	0.24	0.24	0.20	0.20	0.20	0.18	0.04	–0.03
		b	–0.01	–0.01	–0.01	–0.01	–0.01	–0.01	0.00	–0.01	–0.12	–0.12	–0.09	–0.13	–0.13	–0.08	–0.02	–0.03
		α	–0.15	–0.15	–0.19	0.01	–0.10	–0.11	–0.14	–0.01	0.08	0.01	–0.16	0.24	0.24	0.20	–0.21	0.05
		β	–0.05	–0.05	–0.06	–0.01	–0.04	–0.04	–0.05	–0.02	0.09	0.06	0.00	0.13	0.13	0.10	–0.05	0.03
		θ	0.00	–0.01	–0.01	–0.01	–0.01	–0.01	–0.01	–0.01	0.00	0.00	0.00	0.00	0.00	0.00	–0.01	–0.01
		τ	–0.02	–0.01	–0.01	–0.02	–0.02	–0.02	–0.01	–0.02	–0.02	–0.01	–0.01	–0.02	–0.02	–0.02	–0.01	–0.02
	RMSE	a	0.13	0.13	0.12	0.12	0.11	0.11	0.11	0.11	0.39	0.38	**0.31**	0.36	0.33	0.28	0.13	**0.13**
		b	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.17	0.16	**0.13**	0.18	0.14	0.11	**0.06**	0.06
		α	0.15	0.15	0.20	**0.04**	0.11	0.11	0.15	**0.04**	0.12	0.06	**0.17**	0.26	0.40	0.22	0.22	**0.07**
		β	0.05	0.05	0.06	0.02	0.04	0.04	0.05	0.02	0.09	0.06	**0.02**	0.13	0.18	0.11	0.05	**0.03**
		θ	0.30	0.30	0.30	0.29	0.29	0.29	0.29	0.29	0.43	0.42	**0.40**	0.43	0.41	0.40	**0.34**	0.35
		τ	0.11	0.11	0.11	0.10	0.10	0.10	0.10	0.10	0.30	0.29	**0.22**	0.33	0.45	0.39	**0.17**	0.22
40%	bias	a	0.11	0.11	0.09	–0.07	0.03	0.04	0.02	–0.03	0.42	0.42	0.38	0.14	0.35	0.33	0.10	–0.06
		b	–0.03	–0.03	–0.02	–0.02	–0.01	–0.01	–0.01	–0.01	–0.25	–0.23	–0.19	–0.20	–0.22	–0.15	–0.03	–0.02
		α	–0.08	–0.08	–0.13	0.01	–0.02	–0.02	–0.06	–0.02	0.30	0.18	–0.06	0.34	0.72	0.47	–0.18	–0.01
		β	–0.03	–0.03	–0.04	–0.01	–0.02	–0.02	–0.03	–0.02	0.24	0.19	0.08	0.22	0.47	0.28	–0.03	0.01
		θ	0.00	0.00	–0.01	–0.01	–0.01	–0.01	–0.01	–0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	–0.01
		τ	–0.01	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.01	–0.02	–0.02	–0.02	–0.01	–0.02
	RMSE	a	0.17	0.17	0.15	0.14	0.12	0.12	0.11	0.12	0.59	0.57	0.50	**0.32**	0.53	0.45	0.17	**0.16**
		b	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.06	0.32	0.29	**0.23**	0.24	0.27	0.19	0.07	0.06
		α	0.09	0.09	0.14	**0.04**	0.05	0.05	0.07	0.04	0.33	0.21	**0.08**	0.37	0.75	0.49	0.19	**0.05**
		β	0.03	0.03	0.04	0.02	0.02	0.02	0.03	0.02	0.25	0.19	**0.08**	0.22	0.47	0.28	0.03	**0.02**
		θ	0.31	0.31	0.31	0.30	0.30	0.30	0.29	0.29	0.53	0.52	**0.48**	0.50	0.52	0.48	0.39	**0.37**
		τ	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10	0.39	0.36	**0.28**	0.37	0.61	0.49	0.21	**0.18**

注：加粗的表示 RMSE 相对较低的结果。

表　5 情境 2 中各条件下各方法参数估计准确性

π	评价标准	π_i^{non} / 方法	低 OSR	低 CSR	低 CSRI	低 MHM	高 OSR	高 CSR	高 CSRI	高 MHM
20%	bias	a	0.00	0.00	0.00	–0.08	–0.03	–0.02	0.04	–0.08
		b	0.00	0.00	0.00	0.00	–0.07	–0.07	–0.03	–0.07
		α	–0.09	–0.10	–0.14	0.01	0.28	0.19	–0.08	0.37
		β	–0.04	–0.04	–0.05	–0.01	0.15	0.12	0.01	0.18
		θ	–0.01	–0.01	–0.01	–0.01	0.00	0.00	0.00	–0.01
		τ	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02
	RMSE	a	0.11	0.11	0.11	0.14	0.17	0.17	**0.14**	0.18
		b	0.05	0.05	0.05	0.05	0.15	0.14	**0.08**	0.15
		α	0.10	0.10	0.15	**0.04**	0.30	0.21	**0.10**	0.38
		β	0.04	0.04	0.05	0.02	0.15	0.12	**0.02**	0.18
		θ	0.29	0.29	0.29	0.29	0.34	0.34	**0.34**	0.34
		τ	0.10	0.10	0.10	0.10	0.39	0.37	**0.23**	0.41
40%	bias	a	0.02	0.02	0.02	–0.15	–0.07	–0.04	0.07	–0.12
		b	–0.01	–0.01	–0.01	–0.01	–0.15	–0.14	–0.11	–0.15
		α	0.00	0.01	–0.05	0.01	0.57	0.48	0.21	0.65
		β	–0.01	–0.01	–0.02	–0.01	0.38	0.32	0.18	0.41
		θ	–0.01	–0.01	–0.01	–0.01	0.00	0.00	0.00	0.00
		τ	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02	–0.02
	RMSE	a	0.12	0.13	0.12	0.21	0.26	0.24	**0.22**	0.27
		b	0.06	0.06	0.05	0.05	0.27	0.25	**0.17**	0.29
		α	0.05	0.05	0.07	0.04	0.59	0.50	**0.23**	0.67
		β	0.02	0.02	0.03	0.02	0.38	0.32	**0.18**	0.41
		θ	0.30	0.30	0.29	0.29	0.39	0.38	**0.37**	0.39
		τ	0.11	0.11	0.10	0.11	0.51	0.47	**0.36**	0.54

注：加粗的表示 RMSE 相对较低的结果。

从表 5 可以看出，在情境 2 中，π_i^{non}低的条件下，MHM 得到的区分度参数的 RMSE 大于标准化残差系列方法，时间区分度参数 RMSE 小于标准化残差系列方法，此时，MHM 会高估区分度参数，CSRI 会高估时间区分度参数；在π_i^{non}高的条件下，标准化残差系列方法得到的参数估计值 RMSE 整体上都小于 MHM，并且π越大，CSRI 优势越明显。此外，MHM 普遍存在低估时间区分度参数和时间密度参数的问题，在$\pi = 40\%$，π_i^{non}高的条件下，还存在高估区分度参数和难度参数的问题。总的来说，MHM 在情境 2 中的表现比情境 1 差，标准化残差系列方法具有更大的相对优势。

5 研究二：标准化残差系列方法与混合多层模型法比较的实证研究

5.1 数据和设计

研究二使用 James Madison 大学开发的自然界管理测验测试数据，该测验主要测试了学生对与保护环境相关的管理原则、问题和实践应用的了解程度（Pastor et al., 2019）。使用 OSR，CSR，CSRI 和 MHM 对不努力作答进行处理。该测验采用基于网络的方式施测，测量了环境管理原则、问题和实践知识，属于低利害测验。测验长度为 50 题，都是 0/1 计分的选择

题。测试完成后，要求每名被试完成一个关于完成测验努力程度的自陈量表。自陈量表主要包括三个方面内容：（1）认真完成测验重要性评价，分值越高表示重要性程度越高；（2）完成测验努力程度评价，分值越高表示花费的努力程度越高；（3）随机猜测比例，即被试选择自己在完成测验时随机猜测作答题目数量的百分比，分为 4 个选项（0% ~ 5%，6% ~ 25%，26% ~ 50%，大于 50%）。自陈量表的结果能够提供不努力作答识别的效度信息。测试样本为 James Madison 大学 2014 ~ 2015 年秋季和春季学期的学生，共 1532 人。删除了在作答反应、反应时上总缺失比例大于 10%的被试，最终保留 1367 人。应用 OSR，CSR 和 CSRI 分别识别不努力作答并将其替换为缺失，基于 van der Linden（2007）的多层模型估计参数。应用 MHM 同时完成不努力作答的识别和参数估计。各模型先验分布设置与模拟研究相同。

5.2 实证研究结果

实证研究所采用的数据来自于一个低利害测验，并且测验长度较长，预估可能出现较严重的不努力作答。首先，1367 名学生选择随机猜测比例为 0% ~ 5%，6% ~ 25%，26% ~ 50%和大于 50%的学生比例分别为 27.07%，41.92%，22.02%和 9.00%。可以发现，大部分学生不努力作答的严重性程度与模拟研究中不努力作答严重性为低（$\pi_i^{non} \sim U(0, 0.25)$）的情况类似，还有部分学生不努力作答严重性大于这个条件。其次，发现所有被试在所有题目上的对数反应时分布都呈现出双峰分布的特点（Wang, Xu, Shang, & Kuncel, 2018）。因此，数据中可能存在略严重的不努力作答现象。此时，各方法得到的结果差异应当略大，采用 CSRI 或 MHM 可能是较好的选择。

MHM 参数估计不收敛比例为 2.02%，其余方法参数估计全部收敛。后面结果只使用收敛的参数计算。OSR，CSR，CSRI 和 MHM 识别的不努力作答比例分别为 4.69%，5.40%，6.58%和 6.92%，CSRI 和 MHM 识别出的不努力作答比例最大。

5.2.1 识别结果的反应时分布比较

图 7 以一道题目为例，展示了各方法识别出的两种类型作答的对数反应时分布情况。从图中可以看出，OSR，CSR 和 CSRI 识别不努力作答的检验力依次增强。例如，OSR 识别出的两种作答在反应时短的第一个峰的分布中有很大的重合，而 CSRI 和 MHM 几乎能将第一个峰内的所有作答识别为不努力作答。此外，MHM 还会将对数反应时较大的个别作答识别为不努力作答，这是由于该方法假设不努力作答的反应时服从均值和标准差恒定的对数正态

分布，如果估计得到的不努力作答反应时标准差较大（本例中$\hat{\sigma}_c$=1.29），个别识别出的不努力作答可能具有较大的对数反应时。

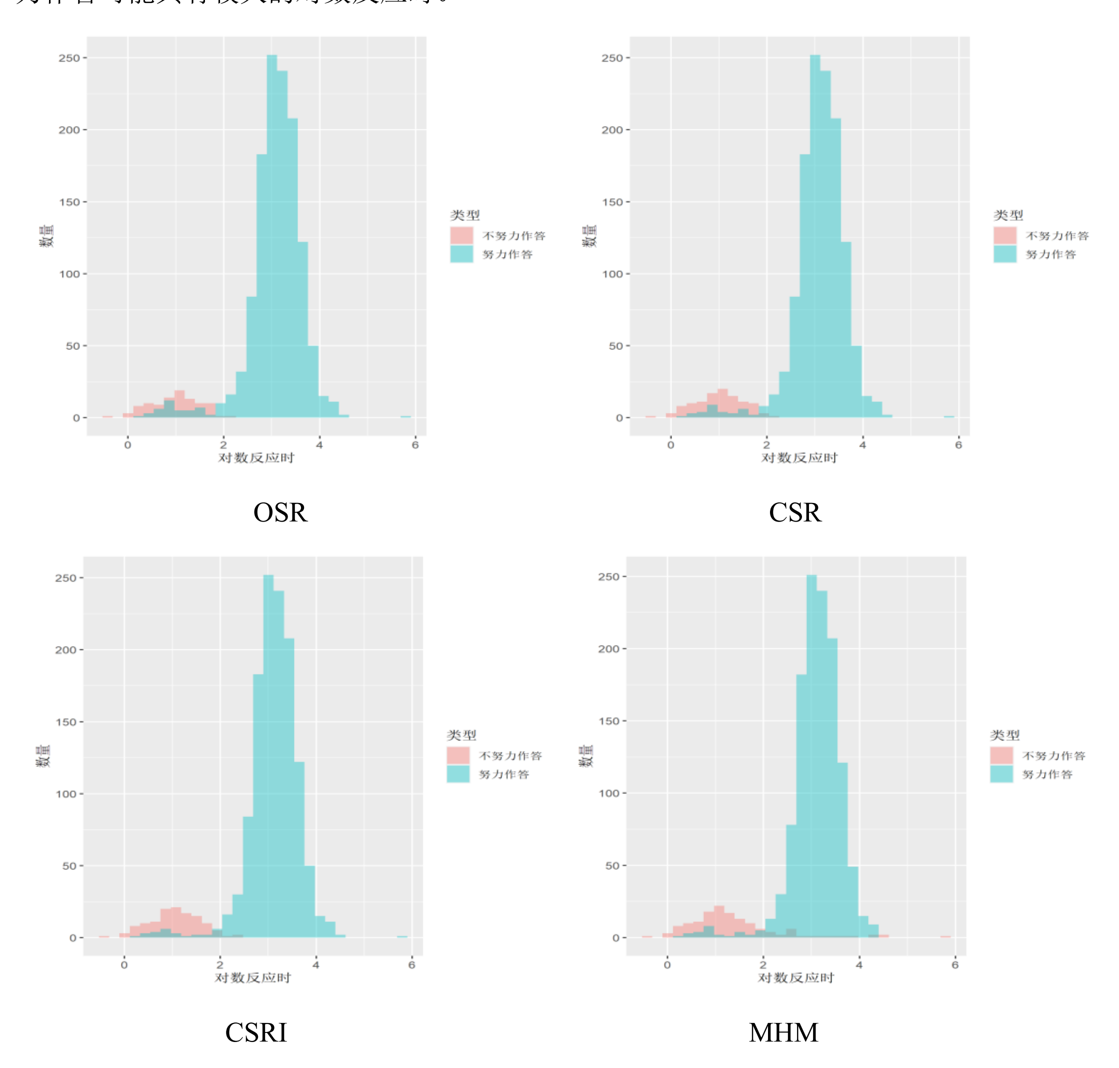


图 1 实证研究各方法识别出的两种作答行为在题目层面反应时分布（以 48 题为例）

5.2.2 识别结果的效度验证

为了对各方法的识别结果进行效度验证，将各方法得到的作答层面识别结果通过 RTE（response time effort, RTE）指标汇总到个体层面，得到每名被试的 RTE 指标。RTE 指标的含义是每个被试努力作答的题目比例。其值越高，说明被试努力作答程度越高。然后，将各方法的 RTE 指标与认真完成测验重要性评价、完成测验努力程度评价的分数求相关。它们与 RTE 指标的相关越高，说明方法会聚效度越高。结果如表 6 所示。从表中可以看出，OSR、CSR、CSRI 和 MHM 与两项评价的相关依次增大，说明其会聚效度依次增大。但是注意到四种方法得到的 RTE 指标与认真完成测验重要性评价相关均较低（低于 0.1）。说明 RTE 指标不同的被试，在认真完成测验的重要性程度评价上只有很小的差异。这与 Pastor 等人（2019）

使用同样数据得到的结果是一致的。

表　6 实证研究不同方法 RTE 指标与认真完成测验重要性评价以及完成测验努力程度评价的相关

RTE	认真完成测验重要性评价	完成测验努力程度评价
OSR	0.055*	0.193**
CSR	0.075**	0.238**
CSRI	0.073**	0.271**
MHM	0.087**	0.288**

注：*表示在 0.05 水平显著，**表示在 0.01 水平显著。

结合 RTE 指标的阈值，可以区分努力作答和不努力作答的被试。基于前人研究建议（Rios et al., 2017）并出于保守考虑，将 RTE 阈值定为 0.8。然后统计两组被试在随机猜测比例四个选项上的选择分布情况。从表 7 中可以看出，努力作答组选择随机猜测比例大于 50%的人数比例最少，不努力作答组选择随机猜测比例为 0-5%的人数比例最少，符合期望的各组特征。此外，CSRI 和 MHM 识别出的努力作答组中选择随机猜测比例大于 50%的人数比例小于另外两种方法，选择随机猜测比例为 0-5%的人数比例大于另外两种方法，但总体来说差异不大。

表　7　实证研究不同组被试在随机猜测比例上选择的人数百分比（%）

方法	分组	0-5%	6-25%	26-50%	大于 50%
OSR	努力作答组	27.96	41.63	21.81	8.60
	不努力作答组	9.23	47.69	26.15	16.92
CSR	努力作答组	28.25	41.94	21.52	8.29
	不努力作答组	10.11	41.57	29.21	19.10
CSRI	努力作答组	28.80	42.00	21.28	7.92
	不努力作答组	8.55	41.03	29.92	20.51
MHM	努力作答组	28.84	41.77	21.12	8.27
	不努力作答组	9.02	43.44	31.15	16.39

为了进一步考察努力作答组和不努力作答组在不同选项上的分布是否存在显著差异，对其进行卡方检验并计算了效应量，卡方检验的效应量采用 Cramer's V 系数（McHugh, 2013）。从表 8 可以看出，使用四种方法识别的努力作答组和不努力作答组，在随机猜测比例上的选择都存在显著差异，并且，使用 CSRI 识别得到的分组在选项上的差异大于 MHM 大于 CSR 大于 OSR。这也可以看作方法会聚效度的另一个证据。总的来说，CSRI 和 MHM 的会聚效度大于其余两种方法。

表　8　实证研究不同组被试在随机猜测比例上选择的卡方检验及效应量结果

方法	卡方值	显著性	效应量
OSR	13.86	0.003	0.20

CSR	23.15	0.000	0.26
CSRI	38.72	0.000	0.34
MHM	29.41	0.000	0.30

5.2.3 估计结果的比较

当数据中存在不努力作答时，使用原始数据会得到有偏差的估计结果，而使用标准化残差系列方法与 MHM 能够在大部分情况下减小参数估计的偏差。为了考察不同方法得到的参数估计结果与基于原始数据得到的参数估计结果之间的差异，首先基于原始数据应用多层模型估计参数作为比较的基线，然后计算不同方法得到的参数估计值与基线参数估计值的相对差异（relative difference, RD）和相对差异均方根（relative root mean square difference, RRMSD），其计算公式与公式（8）（9）类似，区别在于使用基于原始数据估计得到的参数估计值代替原公式中的真值。结果如表 9 所示。

表 9 实证研究不同方法和原始数据参数估计结果比较

参数	RD				RRMSD			
	OSR	CSR	CSRI	MHM	OSR	CSR	CSRI	MHM
a	0.02	0.03	0.06	0.02	0.06	0.08	0.11	0.33
b	0.05	0.07	0.12	0.17	0.10	0.14	0.21	0.28
α	–0.77	–0.84	–0.96	–1.14	0.82	0.90	1.02	1.20
β	–0.11	–0.12	–0.14	–0.02	0.12	0.14	0.16	0.10
θ	0.00	0.00	–0.01	–0.03	0.10	0.11	0.15	0.21
τ	0.00	0.00	0.00	0.09	0.15	0.15	0.20	0.21

注：RD 表示相对差异，RRMSD 表示相对差异均方根。

从表中看出，对于区分度参数，各方法与原始数据得到的估计结果相比都几乎没有差异。对于难度参数，CSRI 和 MHM 得到的估计值小于原始数据的估计结果。对于时间区分度参数，各方法得到的估计结果明显大于原始数据的估计结果，并且差异程度 MHM>CSRI>CSR>OSR。对于时间密度参数，标准化残差系列方法得到的估计结果大于原始数据的估计结果，并且差异程度 CSRI>CSR>OSR，MHM 与原始数据的估计结果相比几乎没有差异。各方法估计得到的被试参数几乎没有相对差异。此外，从整体上看 MHM 和 CSRI 的相对差异均方根也大于其它两种方法。

根据自陈量表关于随机作答比例的报告结果，不努力作答严重性大于模拟研究中π_i^{non}低的条件；又根据 MHM 的识别结果，努力作答和不努力作答的对数反应时差异为 1.529，大于模拟研究中d_{RT}小的条件。可以推测，实证研究的数据较接近模拟研究中π_i^{non}高d_{RT}大的条件。结合表 4 可知，当数据符合 MHM 假设且π_i^{non}高d_{RT}大时，MHM 和 CSRI 表现都优于

CSR，OSR。而实证研究通过效度验证，证明 CSRI 和 MHM 都能够得到较有效的识别结果，并且参数估计值和原始数据估计结果相比差异最大。这与我们对数据中不努力作答情况的预估和方法选择的建议也是一致的。

6 讨论

6.1 方法比较

本研究采用模拟研究和实证研究相结合的方法，对这两类方法进行了比较，得到的结果如下。

从收敛情况来看，标准化残差系列方法由于采用了相对较简单的多层模型，不存在参数估计不收敛的问题。而 MHM 由于在多层模型的基础上加入了两种作答的混合，要多估计 *I**J*+*J*+2 个参数（*I* 表示被试数，*J* 表示题目数，*I**J* 个作答分类潜变量Δ_{ij}，*J* 个不努力作答答对概率参数g_j，2 个不努力作答反应时分布参数μ_c, σ_c）,当样本量很大或题目数很多时，会极大增加待估参数的数量，造成不容易收敛的问题。尤其对于作答分类参数，该问题更严重。除此之外，两类方法的估计速度也不同。笔者通过模拟实验证明，随着测验长度增加，MHM 的耗时明显增加，而其他方法耗时增加相对缓慢。例如，使用处理器为 Intel(R)Core(TM)i7-9700，内存为 32GB 的计算机分析数据，以情境 1 中π为 40%，π_i^{non}高，d_{RT}大的条件为例，当样本量为 2000，测验长度为 10 时，OSR、CSR、CSRI 和 MHM 的耗时分别约为 92 分钟、63 分钟、78 分钟和 240 分钟。其他条件固定，当测验长度增加至 50 题时，四种方法的耗时分别约为 526 分钟、510 分钟、630 分钟和 1160 分钟。同等条件下，即便选用标准化残差系列方法中最复杂的 CSRI 完成识别和参数估计，耗时也仅为 MHM 耗时的约 1/2 以下。因此出于效率的考虑，CSRI 是较好的选择。

从识别情况来看，当数据中含有不努力作答时，CSRI 识别出的不努力作答比例相对较高，并且，其正确识别率（TPR）也基本大于 MHM。在大部分情况下，MHM 的 TPR 甚至小于 OSR 和 CSR，尤其在d_{RT}小的情况下劣势更加明显。这可能是由于此时两种作答反应时差异小，该模型很难根据反应时特征准确区分出这两个类别。例如，有时估计得到的不努力作答反应时分布的均值（μ_c）和标准差（σ_c）都较小（例如，当$\pi = 40\%$，π_i^{non}高d_{RT}小时，30 次重复得到平均估计值$\hat{\mu}_c = -1.14$，$\hat{\sigma}_c = 0.44$），那么基于不努力作答反应时模型的假设，就可能只找出那些反应时极端短的作答而遗漏了大部分反应时相对较长的不努力作

答。而 CSRI 所基于的残差是根据题目参数和速度计算出的，即使实际反应时相对较长，如果速度较慢，仍能够得到较小的残差从而被识别为不努力作答。总的来看，MHM 在识别准确性上依赖于数据产生的模型和两种作答之间的差异，当数据产生的模型符合该方法假设且两种作答反应时差异大时，该方法表现较好，而 CSRI 表现相对稳定。而 CSRI 在大部分条件下错误识别率较高，说明该方法存在超识别问题。在本研究中，该方法 FDR 较大也未造成参数估计误差的增加，这可能是因为参数估计时将作答层面的不努力作答替换为缺失，并未造成样本量明显减少。并且，本研究识别出的不努力作答比例整体不高。根据 Rose（2013）的研究结果，无论缺失机制如何，当整体数据中的缺失比例在 30%以下时，采用忽略的方式得到的参数估计结果是具有稳健性的。因此可以推断，如果 CSRI 识别出的不努力作答比例达到 30%以上，超识别问题可能带来一定程度的参数估计误差，此时选用该方法需要尤其谨慎。

从参数估计情况来看，CSRI 的结果整体上接近或优于 MHM，在d_{RT}小或者产生数据的模型不符合 MHM 假设的情况下，前者的优势更为明显，这与混合多层模型具有强假设的局限是有关的。此外，CSRI 在参数估计方面的缺陷主要是在一些条件下存在一定程度的超识别问题，删除了过多反应时短的作答后，造成反应时分布相对集中，方差变小，时间区分度被高估。此外，由于不努力作答严重性是影响参数估计准确性的重要因素，为进一步探讨不努力作答严重性对各处理方法的影响，基于已有模拟研究，在模拟研究的情境 1 中，固定$\pi = 40\%$，d_{RT}为大，增加了不努力作答严重性的比例的水平，形成无不努力作答，不努力作答严重性低（$\pi_i^{non} \sim U(0, 0.25)$），中（$\pi_i^{non} \sim U(0.25, 0.5)$），高（$\pi_i^{non} \sim U(0.5, 0.75)$）共四个条件。进一步比较各条件下各方法得到的参数估计值 RMSE。结果发现，总的来说，随着不努力作答严重性增加，OSR 和 CSR 得到的参数估计值 RMSE 增加，而 CSRI 和 MHM 得到的 RMSE 基本稳定，它们和另外两种方法的差异逐渐增大。因此，当不努力作答严重性为中或高时，建议选用 CSRI 和 MHM，尤其当严重性为高时这两种方法的优势更强。

6.2 方法总结和建议

标准化残差系列方法和 MHM 从思路上都假设，如果存在异常作答，整个作答反应和反应时都呈现出混合两类模式的特点。但是，它们处理两类作答模式的思路不同。标准化残差系列方法的主要思想是将整个反应时残差分布中极端小的值所对应的作答识别为不努力作答。这类似于假设检验的思路：当整个分布中极端的数值仍属于这个分布时，判断它们不属

于这个分布而犯错的概率是非常小的，因此更有理由相信这些极端的数值属于另一个分布（不努力作答的反应时分布）。然而大量不努力作答会造成参数估计的偏差，进而带来标准化反应时残差的偏差，造成残差不一定服从标准正态分布，严重影响该方法的表现。因此，CSR 在 OSR 基础上，使用了筛选努力作答群体估计题目参数，固定参数并迭代净化这两个策略，在一定程度上提高了识别准确性（Liu & Liu, 2021）。MHM 的基本思想在于用平等的视角对待两类作答模式，将作答反应的正确概率、反应时分布，都视作两类模式的混合。这种思路具有一定灵活性：一是在数据中存在不努力作答的情况下，两类作答分别对各自的模型参数提供信息，不会出现传统模型参数估计误差随不努力作答比例增加而增大的现象；二是从理论上说该模型也能够处理数据中不存在不努力作答的情况，因为此时相当于每个作答的潜类别都相同。但是，该方法包含了强假设，在其不能被满足的情况下结果可能会存在一定偏差。

总的来说，两类方法的特点如表 10 所示。

表 10 研究中比较的四种方法特点小结

比较指标	标准化残差系列方法			MHM
	OSR	CSR	CSRI	
收敛情况	全部收敛	全部收敛	全部收敛	作答分类参数不易收敛
所需时间	短	短	较短	长
正确识别率	不如 CSRI	不如 CSRI	相对最好	不如 CSRI
错误识别率	相对较大	相对较大	相对较大	最低
参数估计准确性	不如 CSRI	不如 CSRI	相对较好，但部分条件下对时间区分度估计误差较大	在数据符合其假设且两种作答反应时差异大的条件下较好
适用情况	不努力作答严重性低	不努力作答严重性低	不努力作答严重性高或中	不努力作答严重性高或中，产生数据符合 MHM 假设，两种作答反应时差异大

根据各方法特点，建议在实际应用中先结合每道题目上被试反应时的分布特征，以及测验是低利害还是高利害测验，预判不努力作答的严重性程度。如果严重性很低甚至可能没有不努力作答，出于效率考虑可以选用最简单的 OSR。如果严重性较低，可以选用标准化残差系列方法或 MHM。如果严重性较高，可以首选 CSRI，但如果应用该方法后发现识别出的不努力作答比例较高（i.e., >30%），可以选用 MHM。

6.3 未来研究展望

本研究也具有一定的局限性，未来研究可以从以下三个方面加以改进。首先，尽管 CSRI 整体表现较好，但仍存在一定缺陷，这可能是由于该方法的超识别问题且直接将不努力作答处理为缺失。未来研究可以考虑对该方法采用更加严格的残差阈值或对反应时模型采用稳健的估计方法(Hong et al., 2021)。稳健的估计方法可以在估计反应时模型参数时，对不努力作答赋予较低的权重，应当能从一定程度上优化时间区分度的估计结果。其次，模拟研究发现，当不努力作答严重性较低时，选用 OSR 和 CSR 更为简便高效。而当不努力作答严重性较高时，CSRI 和 MHM 才表现出较大优势。另外，当数据不符合 MHM 的假设时也不应选择该方法。然而目前，还没有方法能检验实际数据是否符合该模型假设。因此，从提高方法使用效率的角度考虑，未来研究可以基于一些不含强假设方法的初步识别结果，尝试构建一些指标，用于测量整个数据中不努力作答严重程度，或检验数据是否符合 MHM 假设，从而指导实践研究者根据指标选择合适的方法。最后，针对 MHM 估计效率不高的问题，未来研究可以考虑将固定参数的策略应用于 MHM 中，第一步筛选努力作答群体并估计题目参数，第二步将题目参数固定，对其他参数进行条件估计。经初步试验，该策略能将估计时间缩短到原来的一半以下。

7. 结论

本研究得出的主要结论如下：

（1）当不努力作答严重性较低时，标准化残差系列方法和 MHM 在参数估计准确性方面的表现非常接近。

（2）当不努力作答严重性较高时，CSRI 在识别准确性和参数估计准确性方面的表现基本接近或优于 MHM，并且，不存在参数估计收敛的问题，具有更高的效率，在不同情境下具有更好的稳健性，在实际研究中可以作为首选的方法。

（3）MHM 的表现更依赖于数据的具体情况，仅在数据符合其假设且两种作答反应时差异大的条件下有较好表现，并且该方法对于作答分类参数的估计存在不易收敛的问题，识别准确性普遍较低。

参考文献

Borghans, L., & Schils, T. (2012). *The leaning tower of PISA: Decomposing achievement test scores into cognitive and noncognitive components*. The Netherlands: School of Business and Economics, Maastricht University.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI–2 and personality assessment inventory validity indices. *Psychological Assessment, 15*(2), 223–234.

Feinberg, R., & Jurich, D. (2018, April). *Using rapid responses to evaluate test speededness*. Paper presented at the meeting of the National Council of Measurement in Education (NCME), New York, NY.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Hong, M., Rebouas, D. A., & Cheng, Y. (2021). Robust estimation for response time modeling. *Journal of Educational Measurement*. *58*(2), 262–280.

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419.

Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, *80*(4), 775–807.

Liu, Y., & Liu, H. (2021). Detecting non-effortful responses based on a residual method using an iterative purifying approach. *Journal of Educational and Behavioral Statistics*, *online*.

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 261–288.

Matzke, D., Love, J., & Heathcote, A. (2017). A Bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior Research Methods, 49*(1), 267–281.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, *23*(2), 143–149.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205–228.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189–212.

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Retrieved from https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47.

R Development Core Team. (2009). *R: A language and environment for statistical computing* [Computer software Manual]. Vienna, Austria: Retrieved from http://www.Rproject. org (ISBN 3-900051-07-0)

Ranger, J., Wolgast, A., & Kuhn, J. T. (2019). Robust estimation of the hierarchical model for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *72*(1), 83–107.

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated–scores: To filter unmotivated examinees or not?. *International Journal of Testing*, *17*(1), 74–104.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (PhD thesis). Friedrich–Schiller–University, Jena, Germany.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of test–taking effort on a large–scale assessment. *Applied Measurement in Education*, *26*(1), 34–49.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(S1), 365–384.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wang, C., Xu, G., & Shang, Z. (2018). A two–stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501.

Ulitzsch, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86–105.

A Comparison of Standard Residual Methods and a Mixture Hierarchical Model for Detecting Non-effortful Responses

LIU Yue[1], LIU Hongyun[2,3]

([1] *Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu* 610066, *China*)

([2] *Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875, *China*)

([3] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*)

Abstract

Assessment datasets contaminated by non-effortful responses may lead to serious consequences if not handled appropriately. Previous research has proposed two different strategies: *down-weighting* and *accommodating*. Down-weighting tries to limit the influence of aberrant responses on parameter estimation by reducing their weight. The extreme form of *down-weighting* is the detection and removal of irregular responses and response times (RTs). The standard residual-based methods, including the recently developed residual method using an iterative purification process, can be used to detect non-effortful responses in the framework of *down-weighting*. In *accommodating*, on the other hand, one tries to extend a model in order to account for the contaminations directly. This boils down to a mixture hierarchical model (MHM) for responses and RTs. However, to the authors' knowledge, few studies have compared standard residual methods and MHM under different simulation conditions. It is unknown which method should be applied in different situations. Meanwhile, MHM has strong assumptions for different types of responses. It would be valuable to examine the performance of the method when the assumptions are violated. The purpose of this study is to compare standard residual methods and MHM under a fully crossed simulation design. In addition, specific recommendations for their applications are provided.

The simulation study included two scenarios. In simulation scenario I, data were generated under the assumptions of MHM. In simulation scenario II, the assumptions of MHM concerning non-effortful responses and RTs were both violated. Simulation scenario I had three manipulated factors. (1) Non-effort prevalence (π), which was the proportion of individuals with non-effortful responses. It had three levels: 0%, 20% and 40%. (2) Non-effort severity (π_i^{non}), which was the proportion of non-effortful responses for each non-effortful individual. It varied between two levels: low and high. When π_i^{non} was low, π_i^{non} was generated from $U(0, 0.25)$; while when π_i^{non} was

high, π_i^{non} was generated from U (0.5, 0.75), where "U" denoted a uniform distribution. (3) Difference between RTs of non-effortful and effortful responses (d_{RT}). The difference between RTs from two groups, d_{RT}, had two levels, small and large. The logarithm of RTs of non-effortful responses were generated from normal distribution N (μ,0.5^2), where $\mu = -1$ when d_{RT} was small, $\mu = -2$ when d_{RT} was large. For generating the non-effortful responses, we followed Wang, Xu and Shang (2018), with the probability of a correct response g_j setting at 0.25 for all non-effortful responses. In simulation scenario II, only the first two factors were considered. Non-effortful RTs were generated from a uniform distribution with a lower bound of $\exp(-5)$ and upper bound being the 5th percentile of RT on item j with $\tau = 0$. The probability of a correct response for non-effortful responses was dependent on the ability level of each examinee. In all the conditions, sample size was fixed at I = 2,000 and test length was fixed at J = 30. For each condition, 30 replications were generated. For effortful responses, Responses and RTs were simulated from van der Linden's (2007) hierarchical model. Item parameters were generated with $a_j \sim U(1, 2.5)$, $b_j \sim N(0,1)$, $\alpha_j \sim U(1.5, 2.5)$, $\beta_j \sim U(-0.2, 0.2)$. For simulees, the person parameters (θ_i, τ_i) were generated from a bivariate normal distribution with the mean vector of $\boldsymbol{\mu} = (0,0)'$ and the covariance matrix of $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$. Four methods were compared under each condition: the original standard residual method (OSR), conditional estimate standard residual (CSR), conditional estimate with fixed item parameters standard residual method using iterative purifying procedure (CSRI), and MHM. These methods were implemented in R and JAGS using a Bayesian MCMC sampling method for parameter calibration. Finally, these methods were evaluated in terms of convergence rate, detection accuracy and parameter recovery.

The results are presented as following. First of all, MHM suffered from convergence issues, especially for the latent variable indicating non-effortful responses. On the contrary, all the standard residual methods achieved convergence successfully. The convergence issues were more serious in simulation scenario II. Secondly, when all the items were assumed to have effortful responses, the false positive rate (FPR) of MHM was 0. Although the standard residual methods had FPR around 5% (the nominal level), the accuracy of parameter estimates was similar for all these methods. Third, when data were contaminated by non-effortful responses, CSRI had higher true positive rate (TPR) almost in all the conditions. MHM showed lower TPR but lower false discovery rate (FDR),

exhibiting even lower TPR in simulation scenario II. When π_i^{non} was high, CSRI and MHM showed more advantages over the other methods in terms of parameter recovery. However, when π_i^{non} was high and d_{RT} was small, MHM generally had higher RMSE than CSRI. Compared to simulation scenario I, MHM performed worse in simulation scenario II. The only problem CSRI needed to deal with was its overestimation of time discrimination parameter across all the conditions except for when π=40% and d_{RT} was large. In a real data example, all the methods were applied to a dataset collected for program assessment and accountability purposes from undergraduates at a mid-sized southeastern university in USA. Evidences from convergence validity showed that CSRI and MHM might detect non-effortful responses more accurately and obtain more precise parameter estimates for this data.

In conclusion, CSRI generally performed better than the other methods across all the conditions. It is highly recommended to use this method in practice because: (1) It showed acceptable FPR and fairly accurate parameter estimates even when all responses were effortful; (2) It was free of strong assumptions, which meant that it would be robust under various situations; (3) It showed most advantages when π_i^{non} was high in terms of the detection of non-effortful responses and the improvement of the parameter estimation. In order to improve the estimation of time discrimination parameter in CSRI, the robust estimation methods that down-weight flagged response patterns can be used as an alternative to directly removing non-effortful responses (i.e., the method in the current study). MHM can perform well when all its assumptions are met and π_i^{non} is high, d_{RT} is large. However, some parameters have difficulty in convergence under MHM, which will limit its application in practice.

Key words: non-effortful response, standard response time residual, iterative purification, mixture hierarchical model, Bayesian estimation